Over de wiskunde die Google groot maakte

Jan Brandts, Universiteit van Amsterdam

13 januari 2009

Samenvatting

Google vindt in een oogwenk de meest relevante web-bladzijden over een bepaald onderwerp. Omdat het web uit zo'n tien miljard bladzijden bestaat is dit een enorm indrukwekkende prestatie: het lijkt eenvoudiger om een naald in een hooiberg te vinden. De kracht van Google schuilt in de mathematische omschrijving van het begrip belangrijkheid van een web-bladzijde: de zogenaamde PageRank. We zullen laten zien welke ideeen de studenten Page en Brin daar over hadden toen ze Google rond 1997 ontwikkelden. Uiteindelijk bedachten ze een wiskundige vergelijking waarvan de PageRank de oplossing is. We zullen deze PageRank-vergelijking afleiden uit een lijstje van begrijpelijke wensen over PageRank, en ons buigen over de vraag hoe Google deze vervolgens uitrekent. Overigens kunnen we deze vraag niet in zijn geheel beantwoorden: een aantal zaken rondom PageRank wordt door Google strict geheim gehouden. Het Google PageRank probleem wordt door velen gezien als de grootste (matrix-) berekening die ooit is ondernomen.

1 Beknopte geschiedenis

In nog geen tien jaar tijd is Google niet alleen één van de bekendste zoekmachines op het internet geworden, maar is tevens een uiterst succesvol bedrijf, een stuk gereedschap om de effectiviteit van webbladzijdes te analyseren, een werkwoord in (niet alleen) de Engelse taal (to google something), een wereldwijde in detail inzoomende verrekijker (Google Earth), en voor sommigen zelfs een levenswijze. Het woord Googol[1] is in 1920 door de 9-jarige Milton Sirotta bedacht toen zijn vader hem vroeg een passende term te bedenken voor het getal 10^{100}. Het woord Googolplex, waarvan de naam van het hoofdkantoor Googleplex van Google is afgeleid, bedacht hij als naam voor het gigantische getal 10^{googol}. Google werd in 1998 in het leven geroepen door twee studenten van Stanford University in de VS, Larry Page en Sergey Brin. Zij waren één van de eersten die zich realiseerden dat je slim gebruik kunt maken van de structuur van het internet in het proces van informatie vergaren. Informatie vergaren was natuurlijk sinds de geboorte van het World Wide Web in 1989 al hevig veranderd: in plaats van te zoeken middels traditionele gereedschappen zoals kaartenbakken en micro-fishes in bibliotheken en schijven in computers, moest nu ineens het Web worden doorzocht voor informatie. Uiteraard leidde dit tot geheel nieuwe uitdagingen. Immers met zo'n 10 miljard bladzijden is het Web enorm groot. Daarnaast verandert zo'n 40 procent binnen een week van inhoud, en 23 procent zelfs dagelijks en dus is het Web dynamisch. Bovendien zijn er geen getrainde specialisten die verantwoordelijk zijn voor de inhoud en organisatie zoals bij een bibliotheek. Sterker nog, er zijn zogeheten spammers die er een sport van maken om het web opzettelijk te desorganiseren. Het belangrijkste verschil met de traditionele informatiebronnen

[1] Het woord Google ontstond door een spelfout van de investeerders op een cheque aan de oprichters

is echter dat het Web gelinkt is middels hyperlinks. Toch duurde het nog zo'n tien jaar voordat Page en Brin, maar ook Jon Kleinberg van IBM (ook in 1998) zich realiseerden dat het World Wide Web een schoolvoorbeeld is van wat wiskundigen een gerichte graaf noemen en als zodanig behandeld zou moeten worden. Omdat er al vele decennia onderzoek wordt gedaan in de zogeheten grafentheorie, konden de resultaten daarvan meteen worden toegepast op het Web. Dit leidde uiteindelijk tot de ontwikkeling van Google's zogenaamde PageRank, en Jon Kleinberg's minder bekende HITS (Hypertext Induced Topic Search) wat gebruikt wordt in de zoekmachine Teoma.

Opmerking

Het getal googol lijkt erg groot. Enerzijds is dit ook zo. Immers, het aantal elementaire deeltjes in het heelal wordt geschat op zo'n 10^{70}. Anderzijds, als er 70 mensen in de rij staan bij een kassa, is het aantal volgordes waarin ze kunnen staan net even meer dan googol.

2 Het PageRank model

Het succes van Google berust op de wonderbaarlijk goede resultaten die het door Google ontwikkelde PageRank model produceert. Zoals gezegd is de PageRank van een web-bladzijde een getal dat de belangrijkheid van die bladzijde aangeeft. Omdat het Web uit zo'n tien miljard bladzijden bestaat, zijn er tien miljard PageRank getallen. Behalve dat we zullen onderzoeken hoe deze getallen zijn gedefinieerd, is het natuurlijk ook interessant om te kijken naar de problemen die ontstaan doordat al deze getallen met regelmaat uitgerekend moeten worden. Namelijk, omdat het web zo snel van structuur verandert, is het goed mogelijk dat een onbelangrijke bladzijde van vandaag over een maand een belangrijke bladzijde is geworden, en dus zullen die tien miljard getallen met regelmaat moeten worden ververst. Samengevat zullen we ons de volgende vragen stellen.

- Welke criteria worden gebruikt om het belang van web-bladzijden te onderscheiden?
- Hoe zetten we deze criteria om in een wiskundige model?
- Hoe berekenen we op efficiënte wijze de resulterende PageRank getallen?

In de volgende secties zullen we deze vragen proberen te beantwoorden, meestal aan de hand van eenvoudige voorbeelden.

2.1 De intuïtie achter het PageRank model

Het eerste model voor PageRank van Sergey Brin en Larry Page berust op twee hele eenvoudige en logische principes.

Principe 2.1 Een web-bladzijde is belangrijk als ernaar wordt verwezen door andere belangrijke web-bladzijden.

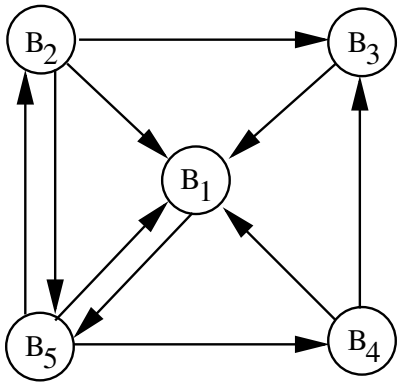
Principe 2.2 Als een web-bladzijde enkel en alleen naar jouw bladzijde verwijst, is dit meer waard dan wanneer deze bladzijde ook naar heel veel ander bladzijdes verwijst.

Beide principes klinken heel natuurlijk, maar lijken niet echt tot concrete getallen voor het belang van een bladzijde te kunnen leiden. Immers, om het belang van een gegeven bladzijde te bepalen aan de hand van rekenregels die op deze principes zijn gebaseerd, moet je

weten hoe belangrijk de bladzijden zijn die naar jouw bladzijde verwijzen. Dit lijkt in een cirkelredenering te verzanden. Verderop zal echter blijken dat dit geen probleem is.

2.2 Een eerste wiskundig model

Laten we een poging doen om bovenstaande twee principes om te zetten in formules die concrete getallen op kunnen gaan leveren. Om de zaken overzichtelijk te houden nemen we aan dat het World Wide Web uit vijf bladzijden bestaat, die naar elkaar verwijzen zoals weergegeven in Afbeelding 3.



Afbeelding 1: Een World Wide Web van vijf bladzijden, met pijlen als hyperlinks.

Met het oog op de beide Principes 2.1 en 2.2 lijkt bladzijde B_1 in dit web een belangrijke bladzijde: alle andere bladzijden verwijzen ernaar. Ook B_5 heeft goede papieren: twee bladzijdes verwijzen ernaar, waaronder de belangrijke B_1. Maar door Principe 2.2 is B_5 misschien wel belangrijker dan B_1, omdat B_1 alleen naar B_5 verwijst, terwijl B_5 behalve naar B_1 ook nog naar twee andere bladzijden verwijst.

Vraag is natuurlijk hoe de heuristieken verwoord in Principes 2.1 en 2.2 omgezet kunnen worden in harde wiskunde. Dat deden Page en Brin als volgt. Laat voor iedere gehele j met $1 \leq j \leq 5$ het symbool P_j staan voor de belangrijkheid van bladzijde B_j. We noemen P_j de PageRank van B_j.

Definitie 2.3 (PageRank) Veronderstel dat bladzijde B_i naar L_i verschillende bladzijdes verwijst, waaronder B_j. Dan draagt B_i een hoeveelheid P_i/L_i bij aan de PageRank van B_j.

In Afbeelding 1 betekent dit dat de PageRank P_1 van bladzijde B_1 wordt berekend middels:

$$P_1 = \frac{P_2}{3} + \frac{P_3}{1} + \frac{P_4}{2} + \frac{P_5}{3}, \tag{1}$$

omdat B_2 naar B_1 verwijst en naar drie bladzijden in totaal, omdat B_3 naar B_1 verwijst en naar één bladzijde in totaal, omdat B_4 naar B_1 verwijst en naar twee bladzijden in totaal, en omdat B_5 naar B_1 verwijst en naar drie bladzijden in totaal. Voor de overige vier bladzijden vinden we op gelijke wijze dat

$$P_2 = \frac{P_5}{3}, \quad P_3 = \frac{P_2}{3} + \frac{P_4}{2}, \quad P_4 = \frac{P_5}{3}, \quad \text{en } P_5 = \frac{P_1}{1} + \frac{P_2}{3}. \tag{2}$$

Het model resulteert dus in vijf lineaire vergelijkingen voor de onbekende getallen $P_1, \ldots, P_5$, die we als volgt kunnen schrijven in matrix-vector vorm:

$$HP = P, \quad \text{waarbij} \quad H = \begin{bmatrix} 0 & \frac{1}{3} & 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 1 & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix} \quad \text{en} \quad P = \begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \end{bmatrix}. \tag{3}$$

Deze vergelijkingen vormen derhalve een speciaal geval van een eigenwaardeprobleem. Er bestaat een oplossing $P \neq 0$ als de matrix H een eigenwaarde gelijk aan één heeft. Middels enig rekenwerk is na te gaan dat dit inderdaad het geval is, en een oplossing P van (3) in gehele getallen is

$$P_1 = 16, \;\; P_2 = 6, \;\; P_3 = 5, \;\; P_4 = 6, \;\; P_5 = 18, \tag{4}$$

ons vermoeden over het belang van B_5 en B_1 bevestigend. Natuurlijk is ook ieder veelvoud van de gegeven oplossing ook een oplossing, maar voor een rangschikking maakt dit niet uit.

Opmerkingen

De rijen en de kolommen van H geven duidelijk weer hoe het web in elkaar zit. De tweede rij van H vertelt je bijvoorbeeld dat B_2 niets ontvangt van bladzijden B_1, B_3 en B_4 en de helft van de PageRank van B_5. De vijfde kolom laat zien dat bladzijde B_5 zijn PageRank gelijkelijk verdeelt over bladzijdes B_1, B_2 en B_4, enzovoorts.

Het world wide web bestaat momenteel uit zo'n tien miljard bladzijden. De matrix in (3) is ongeveer drie bij drie centimeter groot. Op dezelfde schaal is de echte web-matrix H dan zo'n $3.6 \times 10^9\,\mathrm{km}^2$. Met de printer-uitdraai ervan kan je het oppervlak van de aarde zeven maal bedekken.

2.3 De hoeveelheid rekenwerk

Het vinden van de vector P die voldoet aan $HP = P$ kan natuurlijk worden gedaan middels standaard rekenmethoden uit de lineaire algebra, zoals het vegen met rijen en kolommen. Bedenk echter wel dat het benodigde rekenwerk hierbij evenredig is met de derde macht van het aantal web-bladzijden, en dat dit voor tien miljard bladzijden een onmogelijke taak is, zelfs met de snelste supercomputers. In plaats daarvan worden methodes gebruikt die een redelijke benadering, in een klein aantal decimalen, van de exacte oplossing P opleveren, maar dan wel binnen een acceptabele tijd.

Als voorbeeld van zo'n methode bekijken we de volgende zogenaamde dekpunt-iteratie,

$$P^{(k+1)} = HP^{(k)}, \quad \text{met gegeven start-vector } P^{(0)}. \tag{5}$$

Deze iteratie definieert een rij vectoren $P^{(k)}$ waarbij iedere vector in de rij simpelweg H maal de vorige is. Als die rij convergeert voor $k \to \infty$ dan moet dat wel naar een oplossing van $HP = P$ zijn. Een voor de hand liggende keuze voor de start-vector $P^{(0)}$ is de vector met alle entries gelijk aan $1/n$ als de matrix H afmetingen $n \times n$ heeft.

Voor bovenstaand voorbeeld (3) lijkt convergentie inderdaad op te treden. Startend met $P^{(0)}$ gelijk aan de vector met alle entries gelijk aan $1/5$ vinden we voor $P^{(2)}$ in vier decimalen nauwkeurig dat

$$P^{(2)} = (0.3111, 0.0889, 0.0556, 0.0889, 0.4556), \tag{6}$$

en dat weerspiegelt al heel aardig de correcte verhouding gegegeven in (4). Om alle vier decimalen correct te krijgen moesten we wachten tot

$$P^{(30)} = (0.3137, 0.1176, 0.0980, 0.1176, 0.3529). \tag{7}$$

Een voordeel van een dergelijke iteratie is natuurlijk dat je er mee op kunt houden wanneer je wilt en dan toch nog een resultaat overhouden. Bij vegen met rijen en kolommen is dat niet het geval.

Een belangrijke vraag is natuurlijk of bovenstaand voorbeeld typerend is voor de algemene situatie. Heeft iedere op deze manier aan een web gerelateerde matrix H eigenlijk wel een eigenwaarde één? Convergeert de dekpunt-iteratie wel altijd, en kan er iets worden gezegd over de snelheid waarmee dit gebeurt? Antwoord hierop, en aanpassingen van het model die problemen oplossen, geven we in het volgende hoofdstuk.

3 Tekortkomingen

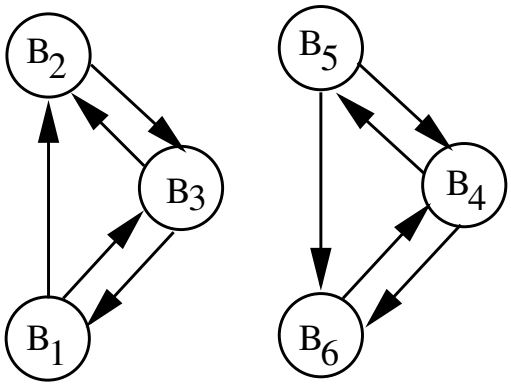
Aan de hand van eenvoudige voorbeelden illustreren we nu wat er allemaal fout kan gaan in het zojuist geïntroduceerde PageRank model. We geven voorbeelden van de volgende problemen:

- Het web is onsamenhangend,
- De dekpunt-iteratie (5) convergeert niet naar de oplossing van de vergelijkingen,
- De PageRank vergelijkingen hebben alleen de oplossing $P_1 = P_2 = \cdots = P_n = 0$.

Dit laatste probleem is het meest serieuze en correspondeert met de opmerking dat H geen eigenwaarde gelijk aan één heeft. Gelukkig kan het model worden aangepast zodat al deze problemen worden opgelost.

3.1 Het web is onsamenhangend

Een eerste tekortkoming van het model uit Hoofdstuk twee is dat het geen uitsluitsel biedt over de situatie waarin het web niet samenhangend is. Hiermee bedoelen we dat er twee of meer verschillende groepen van bladzijden zijn die totaal niet naar elkaar verwijzen, zoals in Afbeelding 2. Voor ieder van beide web-delen kunnen onderling de PageRanks worden uitgerekend, maar omdat veelvouden van PageRanks ook weer PageRanks zijn, is het niet duidelijk hoe de PageRanks van de twee groepen met elkaar moeten worden vergeleken.



Afbeelding 2: Hoe vergelijk je twee volledig afzonderlijke groepen bladzijden?

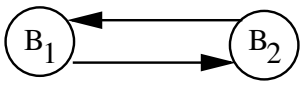
Immers, de PageRanks van het linkerdeel kunnen straffeloos met twee worden vermenigvuldigd terwijl die in het rechterdeel gelijk worden gehouden. Of, in meer wiskundige termen, de eigenruimte van de 6×6 matrix H behorende bij de eigenwaarde één is twee-dimensionaal.

Opmerking

Een kleine losse cluster van web-bladzijden kan onbelangrijk lijken, omdat er niemand naar verwijst. Desondanks kan een web-surfer op één van deze bladzijde aankomen door simpelweg het adres in de adresbalk in te tikken. Behalve het volgen van web-links is ook het intikken van een adres dus een reële mogelijkheid om ergens aan te komen. We zullen dit later in het model verwerken.

3.2 De eenvoudige dekpunt-iteratie convergeert niet

Een tweede tekortkoming is dat de voor de hand liggende dekpunt-iteratie (5) voor de PageRank vergelijking $HP = P$ niet altijd convergeert. Een eenvoudig voorbeeld waaruit dit blijkt is het web met twee bladzijden die naar elkaar verwijzen:



Afbeelding 3: Een web waarvoor de dekpunt-iteratie (5) niet convergeert.

De PageRank vergelijkingen voor P_1 en P_2 zijn natuurlijk

$$P_1 = P_2 \quad \text{en} \quad P_2 = P_1. \tag{8}$$

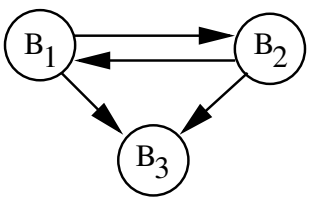
Toch convergeert dekpunt-iteratie (5) niet, als de startvector ongelijk is aan de oplossing. Immers,

$$P^{(k+1)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} P^{(k)}, \quad \text{waarbij } P^{(0)} = \begin{bmatrix} a \\ b \end{bmatrix}, \tag{9}$$

en als $a \neq b$, worden in iedere iteratiestap de vector-entries alleen maar van positie verwisseld. Omdat zoals gezegd de matrix H in de praktijk afmetingen $n \times n$ heeft met n tegen de tien miljard, is (5) één van de weinige praktische mogelijkheden om de oplossing van $HP = P$ in drie a vier decimalen nauwkeurig te kunnen berekenen. En dan nog duurt het met de huidige supercomputers nog een dag of drie! Dus, ondanks dat het wiskundig gezien geen vereiste is voor een goed gedefinieerd PageRank-model, is het praktisch erg wenselijk om op één of andere manier te kunnen garanderen dat (5) met een bepaalde snelheid convergeert.

3.3 De PageRank vergelijkingen hebben geen zinvolle oplossing

Een derde en veel serieuzer probleem is, dat het niet is gegarandeerd dat de PageRank vergelijkingen die worden opgesteld, ook inderdaad een zinvolle oplossing hebben. Hier is een voorbeeld:



Afbeelding 4: Een web zonder zinvolle PageRank oplossing.

De bijbehorende PageRank vergelijkingen zijn de volgende,

$$P_1 = \frac{1}{2}P_2, \quad P_2 = \frac{1}{2}P_1, \quad \text{en} \quad P_3 = \frac{1}{2}P_1 + \frac{1}{2}P_2. \tag{10}$$

De eerste twee vergelijkingen laten zien dat

$$P_1 = \frac{1}{2}P_2 = \frac{1}{2}\left(\frac{1}{2}P_1\right) = \frac{1}{4}P_1, \tag{11}$$

en dus is $P_1 = 0$. Maar dan is ook $P_2 = 0$, en uit de derde vergelijking volgt dat ook $P_3 = 0$. Kortom, de oplossing $P_1 = P_2 = P_3 = 0$ is de enige oplossing van dit stelsel. Of, met andere woorden, de matrix H behorende bij dit web,

$$H = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}, \tag{12}$$

heeft geen eigenwaarde gelijk aan één. Dit is natuurlijk een ongewenste situatie.

Opmerking

De oorzaak voor het ontbreken van een zinvolle PageRank in Afbeelding 4 is dat er een bladzijde is waar weliswaar naar wordt verwezen, maar die zelf nergens naar verwijst. Ongeveer 80 procent van het web bestaat uit dergelijke bladzijden (documenten zoals jpg-, en pdf-files). Dergelijke bladzijden, die in de Engelstalige literatuur dangling nodes heten, kunnen zeer relevante informatie bevatten, dus het zou geen goede oplossing zijn ze gemakshalve maar te negeren.

4 Reparatie van de tekortkomingen

De zojuist geformuleerde drie tekortkomingen van het originele PageRank model kunnen alledrie worden verholpen. We zullen aangeven hoe dit kan worden gedaan, hierbij beginnend met het probleem van de dangling nodes.

4.1 Teleportatie vanuit dangling nodes

Een dangling node is een web-bladzijde die zelf nergens naar verwijst, zoals een jpg- of pdf document. Dergelijke bladzijden kunnen er de oorzaak van zijn dat de PageRank vergelijkingen $HP = P$ alleen de oninteressante oplossing $P = 0$ hebben. Dit is een onvolkomenheid in het oorspronkelijke model, die we als volgt herstellen.

Principe 4.1 (Dangling nodes) Een web-bladzijde die nergens naar verwijst zullen we in het model representeren als een bladzijde die naar alle bladzijden binnen het web, inclusief zichzelf, verwijst.

Dit lijkt de omgekeerde werkelijkheid, maar zo gek is het niet. Immers, je kan in een dangling node geen link aanklikken. Wat je wel kan doen om er weg te komen is een nieuw web-adres intikken in de navigatie-balk van je web-browser. Omdat dit ieder adres op het web kan zijn, is het niet onnatuurlijk om vanuit deze bladzijde pijlen te tekenen naar alle andere bladzijden.

Definitie 4.2 (Teleportatie) De verplaatsing van een dangling node naar een willekeurig ingetypt nieuw web-adres wordt in de literatuur aangeduidt met teleportatie.

Opmerking

De reden om een dangling node ook naar zichzelf te laten wijzen, is dat hij er anders nadeel van zou ondervinden dat hij een dangling node is. De voorgestelde aanpassing lijkt in zekere zin eerlijk.

Wiskundige gezien komt een dangling node overeen met een kolom van H waarin alleen maar nullen staan. Principe 4.1 stelt, dat iedere nul in een dergelijke kolom moet worden vervangen door $1/n$, waarbij n het totaal aantal webbladzijdes is. De matrix H uit (12) wordt op deze manier een matrix S,

$$S = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{3} \end{bmatrix}, \tag{13}$$

wat in zijn algemeenheid kan worden opgeschreven als

$$S = H + \frac{1}{n} a e^T, \tag{14}$$

waarbij $e^T = (1, \ldots, 1)$ en a is de verticale vector met entries $a_j = 1$ als B_j een dangling node is, en $a_j = 0$ als die niet zo is.

Opmerking

Het product ae^T is een zogeheten matrix van rang één. Het beeld van deze matrix zijn de veelvouden van a, immers, een gegeven vector $x \in \mathbb{R}^n$ wordt afgebeeld op a maal de scalar $e^T x \in \mathbb{R}$.

Zelfs al zou de bovenstaande ingreep garanderen dat er altijd een niet-nul oplossing bestaat van de PageRank vergelijkingen, is het duidelijk dat het niet per definitie de andere twee gesignaleerde problemen oplost. Immers, de webs uit Afbeelding 2 en 3 bevatten geen dangling nodes en zullen dus geen wijziging ondergaan ten gevolge van Principe 4.1.

4.2 Een snufje globale teleportatie

Natuurlijk zal een surfer niet alleen als hij in een dangling node aankomt een nieuw web-adres in de navigatie-balk van de web-browser in te tikken. Ook op andere bladzijden zal dit soms worden gedaan, simpelweg omdat de interesse in de huidige bladzijde en de links daarvandaan is verdwenen, of omdat het niet snel genoeg leidt tot de gewenste bestemming.

Principe 4.3 (Globale Teleportatie) Een surfer zal een bepaald deel $1-\alpha$ met $0 < \alpha < 1$ van de tijd niet de links in het web volgen, maar een nieuw adres intikken in de navigatie-balk.

Om dit principe in het model op te nemen doen we het volgende. Het pure teleporteren kunnen we symboliseren middels de zogenaamde teleportatie-matrix T die gedefinieerd is door

$$T = \frac{1}{n} e e^T = \frac{1}{n} \begin{bmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{bmatrix}. \tag{15}$$

Deze matrix staat model voor het (denkbeeldige) web waarin iedere bladzijde naar iedere bladzijde linkt. Je kan met gelijke kans overal naartoe. Kortom, de ultieme teleportatie.

Het echte web zit zo natuurlijk niet in elkaar. Om desalniettemin toch een deel ter grootte $1-\alpha$ uit teleportatie te laten bestaan, en de rest van de tijd ter grootte α het model tot nu toe te volgen, is het uiteindelijke model van Google gebaseerd op de matrix G, genaamd de Google-matrix,

$$G = \alpha S + (1-\alpha)T, \tag{16}$$

die een zogeheten convexe combinatie is van S en T. De parameter α is hierbij dus een getal tussen de nul en de één, waarbij $\alpha = 1$ het oorspronkelijke (en dus niet goed werkende) model voorstelt, en $\alpha = 1$ een (irrealistisch) model waarin de structuur van het web niet meer aanwezig is. Het ligt voor de hand om eens te onderzoeken wat een keuze van α in de buurt van, maar niet gelijk aan één, oplevert, en te zoeken naar een oplossing van

$$GP = P, \tag{17}$$

waarbij G de matrix uit (16) is. Er zal blijken dat het model nu goed gedefinieerd is, in de volgende zin. Bewijzen voor deze uitspraken komen in de afsluitende sectie.

Stelling 4.1 *Kies de parameter α zodanig dat $0 \leq \alpha < 1$. Dan bestaat er een oplossing P van de PageRank vergelijkingen $GP = P$ waarvan iedere entry P_j positief is, en deze entries optellen tot één. Als $GR = R$ dan is R een veelvoud van P. Voor de dekpunt-iteratie*

$$P^{(k+1)} = GP^{(k)} \tag{18}$$

waarbij $P^{(0)}$ een willekeurige startvector is met entries die optellen tot één, geldt dat

$$\|P - P^{(k)}\|_1 \leq \alpha^k \|P - P^{(0)}\|_1. \tag{19}$$

Hier is $\|x\|_1$ gedefinieerd als de som van de absolute waarden van de entries van $x \in \mathbb{R}^n$.

Deze stelling laat zien dat alle gesignaleerde problemen zijn opgelost na het toevoegen van een niet-triviaal snufje globale teleportatie aan het middels locale teleportatie reeds van dangling nodes verloste oorspronkelijke model. Het laat echter ook zien, dat er nog één afweging resteert. Immers, enerzijds zouden we graag α in de buurt van één willen kiezen om zoveel mogelijk de originele webstructuur in het model te betrekken. Anderzijds kan dit een vertraging van de convergentie van de dekpunt-iteratie tot gevolg hebben, zoals blijkt uit (19).

4.3 Voorbeeld

Als illustratief voorbeeld bekijken we het web uit Afbeelding 3, waarvoor de dekpunt-iteratie niet convergeerde. Met de keuze $\alpha = 4/5$ voor de globale teleportatie-parameter vinden we dat

$$G = \frac{4}{5}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + \frac{1}{5}\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 1 & 9 \\ 9 & 1 \end{bmatrix}. \tag{20}$$

Duidelijk is dat voor de dekpunt-iteratie (18) geldt dat

$$P^{(k)} = G^k P^{(0)}. \tag{21}$$

We berekenen de matrix G^k door in te zien dat

$$G = V^{-1}DV, \quad \text{waarbij } D = \begin{bmatrix} -\frac{4}{5} & 0 \\ 0 & 1 \end{bmatrix} \quad \text{en} \quad V = V^{-1} = \frac{1}{2}\sqrt{2}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{22}$$

Hieruit volgt, omdat $G^k = (V^{-1}DV)(V^{-1}DV)\cdots(V^{-1}DV) = V^{-1}D^kV$, dat

$$G^k = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} (-\frac{4}{5})^k & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{23}$$

Startend met een willekeurige vector $P^{(0)} = (q, 1-q)^T$ waarvan de entries optellen tot één, schrijft dit uit tot

$$P^{(k)} = G^k\begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} + \left(-\frac{4}{5}\right)^k (1-2q)\begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = P + \left(-\frac{4}{5}\right)^k (1-2q)\begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}. \tag{24}$$

Hieraan zien we duidelijk dat ook als $q \neq \frac{1}{2}$ de iteratie convergeert naar de oplossing van $GP = P$ met gelijke entries optellend tot één. Ook zien we dat de fout in iedere iteratiestap een multiplicatieve factor α in grootte afneemt.

5 Enkele eenvoudige bewijzen

In het algemeen kan worden bewezen dat in absolute zin de grootste eigenwaarde van G gelijk is aan één, en dat de op één na grootste eigenwaarde van G kleiner dan of gelijk is aan α. De eigenwaarde één is bovendien enkelvoudig: de bijbehorende eigenruimte heeft dimensie één, en ligt bovendien deels in $\mathbb{R}^n_+$, het positieve 2^n-ant (quadrant, octant, etc). De convergentietheorie van de dekpunt-iteratie (5) in algemene vorm resulteert in (19) via een vrij technische analyse gebaseerd op de Jordan normaalvorm van de doorgaans niet-symmetrische matrix G.

Opmerking

Gebruik makend van de zeer specifieke eigenschappen van G kunnen veel bewijzen echter een stuk eenvoudiger. Voor een complete opbouw gebaseerd op oplossingen van $GP = P$ in gehele getallen verwijzen we naar de Webklas Wiskunde [1] van de Universiteit van Amsterdam, die goed geschikt is voor scholieren in de laatste twee jaar van het voorbereidend wetenschappelijk onderwijs.

5.1 Uniciteit

Neem aan dat $0 \leq \alpha < 1$ en dat er een oplossing P bestaat van $GP = P$ waarvan alle entries groter dan of gelijk zijn aan nul. Later zullen we aantonen dat zo'n oplossing inderdaad bestaat, eerst bewijzen we de resterende uitspraken uit Stelling 4.1.

Lemma 5.1 *Als $0 \leq \alpha < 1$ en $GP = P$ en alle entries van P zijn groter dan of gelijk aan nul, dan geldt dat $P = 0$ of dat alle entries van P positief zijn.*

Bewijs. Per definitie van matrix-vector vermenigvuldiging is vector GP de lineaire combinatie van de kolommen van G met als coëfficiënten de entries van P. Als tenminste één van die entries positief is, is wegens het positief zijn van alle entries van G ook GP een vector met positieve entries. Zijn alle entries van P nul dan zijn die van GP dat uiteraard ook. □

Lemma 5.2 *Veronderstel dat $0 \leq \alpha < 1$ en $GP = P$ waarbij P alleen positieve entries heeft. Als R dan voldoet aan $GR = R$, dan is R een veelvoud van P.*

Bewijs. Omdat P alleen positieve entries heeft, bestaat er een $m \in \mathbb{R}$ zodanig dat

$$Q = R + mP \tag{25}$$

ook alleen positieve entries heeft. Laat P_j, Q_j met $j \in \{1, \ldots, n\}$ de entries van P en Q zijn. Dan geldt uiteraard voor alle $j \in \{1, \ldots, n\}$ dat

$$\frac{P_j}{Q_j} \geq q = \min \left\{ \frac{P_i}{Q_i} \;\middle|\; i \in \{1, \ldots, n\} \right\} > 0 \tag{26}$$

en dus ook dat

$$P_j - qQ_j \geq P_j - \frac{P_j}{Q_j} Q_j = 0, \tag{27}$$

waarbij voor ten minste één waarde van j geldt dat $P_j - qQ_j = 0$ omdat het minimum q in (26) wordt aangenomen. De vector $P - qQ$ is dus wegens Lemma 5.1 gelijk aan nul. Dus

$$P = qQ = qR + qmP \tag{28}$$

en dus geldt dat $R = q^{-1}(1 - qm)P$ een veelvoud van P is. □

Hiermee hebben we een deel van Stelling 4.1 bewezen.

5.2 Convergentie van de dekpunt-iteratie

We vervolgen met de studie van de dekpunt-iteratie (18),

$$P^{(k+1)} = GP^{(k)}, \quad \text{met start-vector } P^{(0)}, \tag{29}$$

waarbij we aannemen dat de entries van $P^{(0)}$ optellen tot één. Ook nemen we aan dat de entries van P zelf optellen tot één, wat natuurlijk altijd middels een eenvoudige schaling te realiseren is.

Definieer de rij

$$Y^{(k)} = P - P^{(k)}, \tag{30}$$

dan zien we direct in dat de som van de entries van $Y^{(k)}$ voor iedere k gelijk is aan nul. Dit is een interessante eigenschap, omdat dan ook geldt dat

$$TY^{(k)} = \frac{1}{n} ee^T Y^{(k)} = 0, \tag{31}$$

waarbij T de globale teleportatie-matrix uit (15) is. Immers, $e^T Y^{(k)}$ is nul omdat het precies de som van de entries van $Y^{(k)}$ is.

Definitie 5.3 Voor $Y \in \mathbb{R}^n$ schrijven we Y_+ voor de vector Y met alle negatieve entries van Y vervangen door nul, en Y_- voor de vector Y met alle positieve entries vervangen door nul.

In termen van deze definitie hebben we dus in het bijzonder dat

$$Y = Y_+ + Y_- \quad \text{en} \quad \|Y\|_1 = \|Y_+\|_1 + \|Y_-\|_1, \tag{32}$$

waarbij de definitie van $\|\cdot\|_1$ al gegeven is in Stelling 4.1.

Lemma 5.4 *Voor alle $Y \in \mathbb{R}^n$ geldt dat*

$$\|SY\|_1 \leq \|Y\|_1. \tag{33}$$

Bewijs. Schrijf Y als $Y = Y_+ - Y_-$ volgens Definitie 5.3. Omdat de entries van S niet negatief zijn en kolomsgewijs optellen tot één, geldt dat

$$\|SY_+\|_1 = \|Y_+\|_1 \quad \text{en} \quad \|SY_-\|_1 = \|Y_-\|_1. \tag{34}$$

Maar dan vinden we, middels de driehoeks-ongelijkheid, ook dat

$$\|SY\|_1 = \|SY_+ + SY_-\|_1 \leq \|SY_+\|_1 + \|SY_-\|_1 = \|Y_+\|_1 + \|Y_-\|_1 = \|Y\|_1. \tag{35}$$

Dit bewijst de bewering. □

Lemma 5.5 Er geldt dat

$$\|Y^{(k)}\|_1 \leq \alpha^k \|Y^{(0)}\|_1. \tag{36}$$

Bewijs. Per definitie van $Y^{(k)}$ en $P^{(k)}$ en met behulp van (31) vinden we dat

$$Y^{(k+1)} = P - P^{(k+1)} = G(P - P^{(k)}) = GY^{(k)} = \alpha SY^{(k)}. \tag{37}$$

Hieruit volgt onmiddellijk met behulp van Lemma 5.4 dat

$$\|Y^{(k+1)}\|_1 = \alpha \|SY^{(k)}\|_1 \leq \alpha \|Y^{(k)}\|_1. \tag{38}$$

De bewering volgt nu eenvoudig met inductie. □

Zonder al te ingewikkelde wiskunde te gebruiken zijn we nu dus ook in staat gebleken om de convergentie van de dekpunt-iteratie te bewijzen. Resteert nog aan te tonen dat er inderdaad een oplossing P van $GP = P$ bestaat met positieve entries.

5.3 Existentie

Hier geven we een overtuigend existentie-bewijs gebaseerd op de dekpunt stelling van Brouwer. Deze zeer bekende stelling kent vele versies, maar impliceert bijvoorbeeld dat een continue functie van een gesloten en begrensde deelverzameling $V \subset \mathbb{R}^n$ naar V een dekpunt heeft. De eenvoudigste illustratie hiervan is, dat de grafiek van een continue functie f van het interval $[0, 1]$ naar $[0, 1]$ de diagonaal van $[0, 1] \times [0, 1]$ wel moet snijden in een zeker punt $z \in [0, 1]$ waar dus blijkbaar geldt dat $f(z) = z$.

Beschouw daarom nu de verzameling

$$V = \{x \in \mathbb{R}^n \mid \forall j \in \{1, \ldots, n\} : x_j \geq 0, \quad \text{en} \quad \sum_{k=1}^{n} x_k = 1\}. \tag{39}$$

In $\mathbb{R}^2$ is dit bijvoorbeeld het lijnstuk tussen de punten $(1, 0)$ en $(0, 1)$. In $\mathbb{R}^3$ is het de driehoek tussen de punten $(1, 0, 0), (0, 1, 0)$ en $(0, 0, 1)$. In het algemeen is het de $(n-1)$-dimensionale simplex tussen de toppen van de n canonieke basis-vectoren van de $\mathbb{R}^n$. Duidelijk is dat V een gesloten en begrensd deel van $\mathbb{R}^n$ is. Bovendien geldt, omdat de kolommen van G optellen tot één, dat met $e^T = (1, \ldots, 1)$,

$$e^T G = e^T, \quad \text{en dus ook dat} \quad e^T G x = e^T x. \tag{40}$$

Omdat $e^T Gx$ de som van de entries van Gx is, is deze som gelijk aan de som $e^T x$ van de entries van x. Ook is duidelijk dat als x geen negatieve entries heeft, ook Gx geen negatieve entires heeft, immers, alle entries van G zijn positief. We concluderen dat $G : V \to V$, en omdat G lineair is, is G ook zeker continu.

De dekpuntstelling van Brouwer garandeert nu het bestaat van een $P \in V$ met $GP = P$. Omdat $0 \notin V$ zegt Lemma 5.1 dat deze P alleen maar positieve entries heeft.

Opmerking

Er zijn diverse manieren om het bestaat van een P met positieve entries aan te tonen waarvoor $GP = P$. De meeste vereisen toch enige wiskundige voorkennis, zoals het bewijs dat we hierboven hebben geven. Een uitzondering is een methode die het bestaan van zo'n P met zelfs gehele positieve entries bewijst. Hiervoor volstaat de aanname dat de parameter α rationaal is. Het bewijs vergt echter een hele andere kijk op het Google PageRank probleem, en we verwijzen de geïnteresseerde lezer daarom ook naar de Webklas Wiskunde [1] van de Universiteit van Amsterdam.

6 Slotopmerkingen

We hebben gezien hoe de oprichters van Google een in eerste instantie eenvoudig model opstelden voor het belang, de PageRank, van web-bladzijden. Om te garanderen dat er een unieke PageRank vector bestaat met positieve entries die optellen tot één moest locale teleportatie vanuit dangling nodes worden toegevoegd, en een snufje globale teleportatie. Om van de uiteindelijke Google matrix G de PageRank vector P met $GP = P$ ook daadwerkelijk uit te rekenen, moet noodgedwongen gebruik worden gemaakt van benaderende, iteratieve methoden zoals de in dit verhaal bestudeerde simpele dekpunt-iteratie.

We hebben niet expliciet opgemerkt dat in het resulterende model de problemen ten gevolge van het onsamenhangend zijn van het web, zijn opgelost. We dagen de lezer uit om gewapend met het volledige model, de PageRanks van het web uit Afbeelding 2 dan wel expliciet uit te rekenen, of middels een iteratie te benaderen!

Dankwoord

De auteur is Fokko van de Bult zeer erkentelijk voor zijn medewerking in de Webklas Wiskunde [1] van de UvA en de hieruit volgende grote bijdrage aan de vereenvoudigde bewijzen, zoals het bewijs van de convergentie van de dekpunt-iteratie in 5.2, dat noch eigenwaarden, noch Jordan normaalvormen vereist. Daarnaast is het boek [2] een enorme inspiratie geweest voor verdere verdieping in het PageRank probleem.

Referenties

[1] J.H. Brandts en F.J. van de Bult (2008) UvA Webklas Wiskunde: de PageRank van Google in 12 uur. http://staff.science.uva.nl/∼brandts

[2] A.N. Langville and C.D. Meyer (2006) Google's PageRank and Beyond: the science of search engine rankings. *Princeton University Press*, Princeton and Oxford.